Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 39 (2006) 626–636
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Terminology model discovery using natural language processing and visualization techniques

Li Zhou [a,*,1], Ying Tao [a,*,1], James J. Cimino [a], Elizabeth S. Chen [a], Hongfang Liu [b], Yves A. Lussier [a], George Hripcsak [a], Carol Friedman [a]

[a] *Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[b] *Department of Information Systems, University of Maryland, Baltimore, MD, USA*

Received 29 June 2005
Available online 29 November 2005

Abstract

Medical terminologies are important for unambiguous encoding and exchange of clinical information. The traditional manual method of developing terminology models is time-consuming and limited in the number of phrases that a human developer can examine. In this paper, we present an automated method for developing medical terminology models based on natural language processing (NLP) and information visualization techniques. Surgical pathology reports were selected as the testing corpus for developing a pathology procedure terminology model. The use of a general NLP processor for the medical domain, MedLEE, provides an automated method for acquiring semantic structures from a free text corpus and sheds light on a new high-throughput method of medical terminology model development. The use of an information visualization technique supports the summarization and visualization of the large quantity of semantic structures generated from medical documents. We believe that a general method based on NLP and information visualization will facilitate the modeling of medical terminologies.

Keywords: Terminology model; Natural language processing; Information visualization

1. Introduction

It is widely recognized that unambiguous encoding of clinical information is crucial for many medical informatics systems, such as clinical information exchange, clinical decision support, information retrieval, and data mining analysis [1,2]. In order to meet the needs of these systems, terminology researchers in biomedical informatics have proposed possible solutions to address issues about the structure and content of terminologies. One of the efforts that has been pursued is to generate standard terminology models in the medical domain [3,4]. Unlike the development of terminologies that involves the discovery of terms, terminology model development is concerned with discovering the underlying model for these terms. In general, there are two kinds of terminology models in biomedical informatics that researchers usually refer to. One is an overall model for an entire controlled vocabulary; e.g., SNOMED-CT [5] uses a formal logic to define concepts and organizes them into hierarchies. The other is a formal representation (or semantic structure) for a coherent class of terms in specific domains, such as radiology [6–8], nursing [9–11], anatomy [12–16], and surgical procedures [17–20], to reflect the minutia of medical concepts that are used daily by clinicians. In these models, elementary concepts are identified, and complex medical concepts are expressed in structured formats, in which the semantic relations of the elementary concepts are well-defined [3,21]. As a result, the clear semantic relations between elementary concepts eliminate lexical variants such as multiple ways of composing terms that have the same meaning, and therefore elim-

* Corresponding authors. Fax: +1 212 342 1647.
E-mail addresses: li.zhou@dbmi.columbia.edu (L. Zhou), ying.tao@dbmi.columbia.edu (Y. Tao).
[1] These two authors have contributed equally to the work.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.10.006

inate possible redundancy and inconsistency. In addition, information structured by a formal model is more computationally tractable.

Despite their advantages, medical terminology models are initially expensive to build. Most of the development is conducted manually by domain experts and linguists. Collecting, summarizing, and analyzing the original terms in biomedical text and then obtaining a consensus model often involves the participation of many groups and many iterations in a collaborative manner [7]. Furthermore, due to the limited number of concepts that can be examined manually, the generated model may not be adequately representative. Low frequency of occurrence but nevertheless significant concepts are easily overlooked, and this may affect completeness [22]. Thus, automated and efficient methods that can help to discover terminology models are strongly needed [23].

In this paper, we introduce an automated method that can facilitate the discovery of medical terminology models. We used pathology procedure concepts as our test domain. This research has two purposes. One is to investigate the feasibility of using natural language processing (NLP) to extract the language patterns of medical concepts from the corpus of original medical text in a highly specific medical domain as a way of helping to educe a terminology model. The medical corpus may contain a large number of concepts; thus it would be difficult to manually summarize the considerable volumes of relations among them. Therefore, the second purpose of this research is to build a tool to visualize and summarize the discovered concept patterns so that a terminology model can be elicited based on a large number of concepts from a corpus.

The organization of this paper is as follows. First, in the background section, a brief review of the traditional procedures for developing medical terminology models is given. Then, an NLP system, MedLEE, used to parse the syntactic structure of pathology report concepts in this research, is introduced. In the methods section, we describe in detail the methodology of extracting the semantic structure of pathology procedure concepts from the corpus using MedLEE. We then introduce a visualization program that summarizes the semantic structure of pathology procedure concepts and facilitates the extraction of a semantic model. Results of our methods and discussion of the advantages and limitations follow the methods section.

2. Background

2.1. Manual approaches to model development

The manual approach is the basic method taken by many developers of medical concept models [6,7,23–28]. This approach requires the participation of human experts who have specific knowledge in the domain during the processes of concept modeling. The procedure usually includes examination of a corpus of noun phrases that are extracted from certain types of clinical reports in natural language. The task of human experts is to go through these noun phrases and construct the terminology model for most of the concepts conveyed in the noun phrases. The process involves summarizing and creating concept types and relations between certain concept types. Another important task of human experts is to evaluate alternative models and decide on a final consensus model. Manual approaches have the advantage of high precision because of the involvement of human experts. Evaluation and selection of alternative models can only be done on the basis of human knowledge. However, the drawback comes from the limited corpus that can be analyzed by human experts; hence the generated model may not be adequately representative. Also low frequency of occurrence examples may be easily omitted, which may affect the general acceptance of the generated models. Another disadvantage is that manual approaches are labor-intensive and time-consuming. Acquiring a consensus model often requires a collaboration of many groups. For example, a project to develop a compositional terminology model for nursing orders took a period of over 3 months [24], which we believe to be usual for manual approaches.

2.2. Automated approaches to model development

Cimino [29] reviewed different terminology tools including terminology browsers that allow users to look-up terms and navigate the structure of a terminology, terminology editors for the maintenance of terminologies, and terminology servers that support the integration of clinical applications. Compared to these efforts, however, there are fewer tools for terminology and terminology model construction. Several automated approaches have been attempted during certain stages of the process. A method was used by Baud et al. [30] to obtain a semantic model mainly for pathological processes. Their approach combined several NLP techniques, including tokenization, rule-based syntactic parsing, semantic tagging, and a rule base for generating semantic relations. However, no visualization method for summarizing the generated models was used. Another attempt was the Cassandra system developed in Europe for assisting the manual modeling process [31]. The Cassandra system and our system are similar in that both employ an automatically generated format to represent the structures of concepts and to assist terminology modeling. However, instead of using XML as the structural format for semantics, Cassandra uses a set of ad hoc tags as the mediator between linguistic representation and formal model. Moreover, our system has a visualization tool for summarizing the semantic structures.

Friedman et al. [32,33] developed a tool, called DynTreeViewer, for vocabulary development and maintenance based on an NLP system, MedLEE. Since our approach also uses MedLEE and some concepts employed by DynTreeViewer, in this section, we will introduce some background of MedLEE and DynTreeViewer.

2.2.1. MedLEE

MedLEE, the Medical Language Extraction and Encoding System, was developed as a general natural language extraction and encoding system within the medical domain [34–39]. It can handle reports in areas, such as radiology, mammography, pathology, echocardiography, electrocardiography, as well as discharge summaries. This system has been widely applied to medical research and practical applications, including automated encoding of clinical documents [38,40], medical error detection [41,42], information retrieval [43,44], clinical research [45–47], and medical terminologies [48,49]. MedLEE has a knowledge component for clinical text (the lexicon) that is used to classify single-word and multi-word phrases and to specify their canonical forms. Using the decompositional option of parsing, multi-word phrases are considered as compositional and each word is considered independently, so that MedLEE structures the individual element words based on the conceptual relations between them. The current study mainly uses the decompositional option of MedLEE.

An important structured output format of MedLEE is XML [37], which carries semantic information about the data while retaining the original contents. The XML document can be viewed as a tree in which each tag is a node. The tree is presented so that more general information (e.g., findings, problems, and procedures) is displayed at the top level and more specific modifying information (e.g., modifiers of related findings) is displayed at the lower levels. Another important component of MedLEE is automated UMLS [50] (Unified Medical Language System) encoding. Structured output generated by MedLEE, consisting of findings and their modifiers, is encoded by the most specific UMLS code. Details of this feature were presented in [38]. Fig. 1 shows a simplified XML output of a pathology term *right dorsal bladder neck biopsy*. The example in Fig. 1 focuses on *biopsy* whose semantic type is *procedure* in the top level of the XML tree. *Biopsy* has a modifier whose semantic type is *bodyloc* with a value *bladder*, which in turn has a modifier whose semantic type is *region* with a value *neck* which again has a modifier whose semantic type is *position* with a value *dorsal*. Finally, *dorsal* has a modifier whose semantic type is *region* with a value *right*. Importantly, MedLEE attempts to assign UMLS codes for an entire term as well as each elementary term within the entire term. For example, in Fig. 1, two UMLS codes are assigned to the elementary term "biopsy" (C0005558 and C0184921). In contrast, code C0194379 represents a combination of "biopsy" and "bladder." No code was found for the entire term.

2.2.2. DynTreeViewer

DynTreeViewer builds a tree that is obtained in two steps: (a) by parsing and encoding a large corpus of text reports in a particular domain to obtain XML structured output, and (b) by combining the XML output for the complete corpus that was generated by the parser. Once a merged tree is constructed, a graphical user interface allows users to see term frequency, relations of terms to other terms, the compositional components of terms, and correspondences to UMLS codes, via a flexible XML-based tree structure. The tree can be viewed, dynamically manipulated, and edited using the interface. The rich structural information of terms comes from the processing of a large collection of patient reports using MedLEE. However, the goal of DynTreeViewer is mainly for vocabulary development and maintenance at the level of individual terms, not for visualizing and summarizing vocabulary models as a whole for a domain. To our best knowledge, no visualization methods have been used to summarize the large number of models generated by automated approaches, such as NLP.

3. Methods

In this research, we selected surgical pathology procedures as the initial domain. An overview of our method

```
<structured form = "xml">
  <procedure v = "biopsy" code = "UMLS:C0005558^biopsy|UMLS:C0184921^excision biopsy" idref = "p10">
      <bodyloc v = "bladder" code = "UMLS:C0005682^bladder|UMLS:C1281573^entire bladder" idref = "p6">
         <region v = "neck" idref = "p8"></region>
         <position v = "dorsal" idref = "p4">
             <region v = "right" idref = "p2"></region>
         </position>
         <code v = "UMLS:C0227716^neck of urinary bladder" idref = "p6 p8"></code>
      </bodyloc>
      <code v = "UMLS:C0194379^biopsy of bladder" idref = "p6 p10"></code>
   </procedure>
</structured>
<tt> <sent id = "s1.1.1"><phr id = "p2">right</phr> <phr id = "p4">dorsal</phr> <phr id = p6">bladder</phr> <phr id
= p8">neck</phr> <phr id = "p10">biopsy</phr>.</sent></tt></section>
```

Fig. 1. An example of XML output of MedLEE for the term *right dorsal bladder neck biopsy*. Tags that are not related to this study were removed for simplicity. The output contains two parts. The *structured* part displays structured findings in XML format and the *tt* part presents the original text. The structured information and the original terms are linked to each other by attribute "idref" and tag "phr." In each tag, the name of the semantic type is followed by a value attribute represented as "v," a UMLS encoding attribute represented as "code," and a phrase identification attribute represented as "idref."

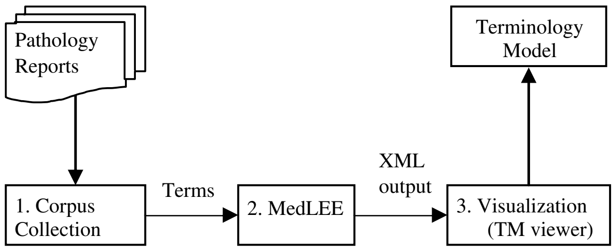

Fig. 2. Overview of the processing steps of the method: (1) target terms were extracted from a large collection of pathology reports; (2) a NLP system (MedLEE) was used to parse the terms and generate XML output; (3) a JAVA Graphical User Interface (GUI) tool, called TMviewer, merges XML output from the XML structures of the individual terms and shows the summarized semantic structures, UMLS encoding and related statistics of the parsed terms; and (4) terminology model was built based on previous steps.

is shown in Fig. 2. In the first step, corpus collection, pathology procedure terms were extracted from the specimen sections of surgical pathology reports. In the second step, the MedLEE system was used to parse the terms and generate coded XML output, representing the compositional structure of the terms. The coding system used was the UMLS. The third step involved using a JAVA Graphical User Interface (GUI) tool we created, called TMviewer, which shows the summarized semantic structures, and related statistics of the parsed terms, along with corresponding UMLS codes and coverage. Users can browse and interact with TMviewer to conceptualize the terminology model based on the information provided by TMviewer.

3.1. Term identification

The corpus consisted of narrative surgical pathology reports recorded over a 10-year period from 1991 to 2000 at Columbia University Medical Center. The first step of the process involved identifying candidate terms. Each report contained general information such as patient name and date, procedures and specimens, as well as descriptions and interpretations of them. Each type of information was stored in a separate section, for example, procedures and specimens were stored in the "Specimen" section. A computer program was used to collect all the pathology procedure terms in this section. The specimen sections contained non-semantic unique identifiers, like "CG91-15734," so these symbols were removed. For each unique term, the frequency of occurrence was computed. This produced a list of terms with the frequency of occurrence of each, such as "Papanicolaou Smear|230312."

3.2. MedLEE parsing

Because the purpose of this study is to look at semantic relations among components of terms, we used MedLEE's decompositional parsing option to extract these relations. The decompositional parsing option ignores multi-word lexical items, such as "breast biopsy." For example, the term "left breast biopsy" will be decomposed into "procedure: biopsy" which is modified by "body location: breast" which is further modified by "region: left." The lexical entry for "breast biopsy" is ignored in the decompositional option of MedLEE. The generated semantic structure for each phrase was encoded in an XML file. The frequency of occurrence in the original term list was copied to each node in the XML file. Simplified versions of three sample XML files are shown in Figs. 3A–C. The terms and their frequencies of occurrence are shown in the first line of each figure, and the processed output follows.

3.3. TMviewer

The development of the visualization tool required two steps: merging of semantic structures from XML files and building a user interface.

3.3.1. Merging semantic structures from XML files

Before the summarized semantic structure of the corpus can be presented, a merging process is needed to combine all the XML files, whose structures represent complex terms referring to pathology procedures, such as "left breast biopsy." This process is handled by a built-in algorithm within TMviewer. The merging is similar to the process used by DynTreeViewer, but the focus is on merging semantic types rather than individual terms. In order to merge two semantic structures, top level semantic types are compared first. If the two structures share the same semantic type, the top semantic types will be merged and their frequencies of occurrence summed. For example, in Fig. 3, for "liver biopsy" and "left breast biopsy," the top semantic types are both "procedure." Thus, the semantic type "procedure" will be the top semantic type after merging. Its frequency of occurrence will be the sum of the two terms (i.e., $564 + 365$). If two terms' top semantic types are different, they will be kept intact after the merging process. Similar merging processes happen in each level of the semantic structures of two terms. Fig. 4 shows the summarization of the semantic structures and frequencies of occurrence after merging the three terms shown in Figs. 3A–C. In this way, a tree structure of semantic types and associated frequency of occurrence information was constructed for the entire corpus.

3.3.2. User interface

After the semantic types were merged together according to their hierarchical positions in the XML tree, a visualization interface in TMviewer was used to visualize the semantic structures of the corpus. Fig. 5 shows a screen snapshot of the interface. The left pane of TMviewer visualizes the tree of the semantic structure of pathology procedure terms. When a node in the tree that represents a semantic type is selected, detailed information about the elementary concepts under that semantic type or about the compositional concepts across multiple semantic types

A **liver biopsy | 564**

```
<procedure v = “biopsy” code = "UMLS:C0005558^biopsy|UMLS:C0184921^excision biopsy” idref = "p4"
  occurrence=564>
    <bodyloc v= “liver” code = “UMLS:C0023884^liver|UMLS:C0205054^hepatic|UMLS:C1278929^entire
      liver" idref = "p2" occurrence=564 >
    </bodyloc>
    <code v = "UMLS:C0193388^biopsy of liver" idref = "p2 p4"></code>
</procedure>
```

B **left breast biopsy | 355**

```
<procedure v = "biopsy" code = "UMLS:C0005558^biopsy|UMLS:C0184921^excision biopsy" idref = "p6"
  occurrence=355>
    <bodyloc v = "breast" code = "UMLS:C0006141^breast|UMLS:C1268990^entire breast" idref = "p4"
      occurrence=355>
        <region v = "left" idref = "p2" occurrence=355></region>
        <code v = "UMLS:C0222601^left breast structure" idref = "p2 p4"></code>
    </bodyloc>
    <code v = "UMLS:C0405348^excisional biopsy of breast" idref= "p4 p6"></code>
    <code v = "UMLS:C0405352^biopsy of breast" idref = "p4 p6"></code>
</procedure>
```

C **breast, right, needle core biopsy | 343**

```
<procedure v = "biopsy" code = "UMLS:C0005558^biopsy|UMLS:C0184921^excision biopsy"idref = "p12"
  occurrence=343>
    <bodyloc v = "breast" c ode = "UMLS:C0006141^breast|UMLS:C1268990^entire breast" idref= "p2"
      occurrence=343>
        <region v = "right" occurrence=343></region>
        <code v = "UMLS:C0222600^right breast structure" idref = "p2 p5"></code>
    </bodyloc>
    <descriptor v= "needle core" idref= “p8” occurrence=343></descriptor>
    <code v = "UMLS:C0405348^excisional biopsy of breast" idref = "p2 p12"></code>
    <code v = "UMLS:C0405352^biopsy of breast" idref = "p2 p12">
</procedure>
```

Fig. 3. Example XML outputs of MedLEE. The terms and their occurrences are shown in the first line, and the processed structured output followed. The elements are semantic types (e.g., *procedure*). Each has attributes of *v* (value), and *idref* (sentence position), and their values. UMLS Concept Unique Identifiers (CUIs) are listed for both atomic and compositional terms if available.

```
<procedure occurrence=1262>
  <bodyloc occurrence=1262>
    <region occurrence=698></region>
  </bodyloc>
  <descriptor occurrence=343></descriptor>
</procedure>
```

Fig. 4. Summarization of semantic structures from Figs. 3A–C with summed occurrences. The semantic type at the top level of the merged XML tree of these three terms is *procedure* with an occurrence of 1262 (564 + 355 + 343). The semantic types in the second level of the XML tree are *bodyloc* and *descriptor*, and their occurrences after merging were 1262 (564 + 355 + 343) and 343, respectively. The semantic type in the third level of the XML tree is *region* with occurrence of 698 (355 + 343).

along the paths is displayed in the upper right pane. For example, in Fig. 5, the user clicked on “procedure,” and was shown procedures such as biopsy, excision, curettage, etc. in the upper right pane. When clicking on each elementary or compositional concept in this list, the original text that generated these concepts is displayed in the lower right text area. For example, the user clicked on “excision,” and was shown all related original terms.

The merged tree in the left pane is displayed using an expandable tree so that branches of the tree can collapse into single nodes if the tree has too many branches to be displayed simultaneously. For example, Fig. 6 shows the completely expanded subtree of procedure, whereas in Fig. 5 the procedure subtree has not been expanded. To hide those semantic structures of low frequency of occurrence, TMviewer has a function called “threshold” for pruning the tree by showing only those nodes which occur above a designated threshold. In this way, TMviewer can help users to get a quick overview of the merged tree structure. For example, in Fig. 5 the threshold is set at zero, showing all nodes. In contrast, the threshold in Figs. 7 and 8 was 10,000, thus the tree shows only the nodes whose frequencies of occurrence are equal to or greater than 10,000, and the tree is expanded. Additionally, the semantic types can be sorted by names and frequencies of occurrence in descending or ascending order based on the users' preferences. For example, in Fig. 5 the tree is sorted by frequency of occurrence of each semantic type in descending order, and the semantic type “procedure” in the top level

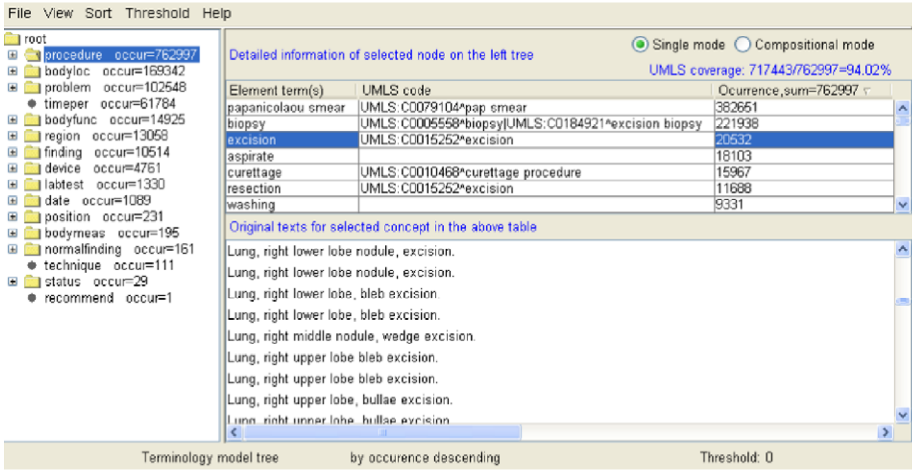

Fig. 5. A screen snapshot of TMviewer (Threshold = 0). Left pane of TMviewer visualizes the tree of the semantic structure of pathology procedure terms. It is sorted by occurrence of each semantic type in descending order, and the semantic type "procedure" in the top level of a tree is selected. The right pane displays detailed information for the elementary concepts that belong to the semantic type selected in the left pane. The first column "Element term(s)" in the table shows the original terms whose semantic type is "procedure," the second column "UMLS code" shows the matched UMLS codes if available, and the third column shows the occurrence of the terms in the corpus. The lower screen of the right pane displays the original text of the selected term. The Single Mode is selected and the table is sorted by occurrence, so that the terms with or without UMLS codes are grouped separately. UMLS coverage for the terms with semantic type of "procedure" is 94.02%.

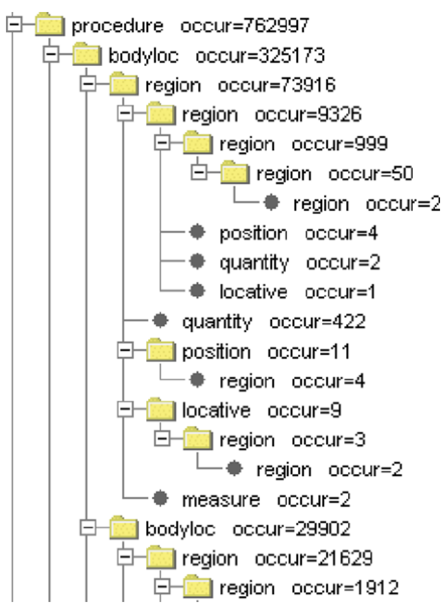

Fig. 6. A subset of the expanded semantic tree of procedure.

of a tree is selected. The symbol "•" in the tree indicates that the node is a terminal node and has no descendants with the designated pruning threshold.

While visualizing the semantic tree structure, TMviewer also provides the function of exploring the semantic tree by showing detailed information. There are two modes for showing the detailed information, Single Mode and Compositional Mode (offered in the upper right pane). Single Mode is for displaying detailed information about elementary concepts belonging to a single semantic type. Compositional Mode is for displaying detailed information about compositional concepts whose elementary concepts are from multiple semantic types.

In Single Mode (e.g., Figs. 5 and 7), elementary concepts, UMLS codes, and their frequencies of occurrence for just the selected semantic type are displayed in a table on the upper right pane of screen. The three columns of the table are: (1) primary terms corresponding to the structured output extracted by MedLEE, (2) their UMLS codes, (3) and their frequency of occurrence within the corpus. Clicking on each column title will sort the table according to that column by ascending and descending order alternately. For example, Figs. 5 and 7 are sorted by frequency of occurrence and Fig. 8 is sorted by UMLS code. A good use of this sorting is to see what the most prevalent terms are in the domain, or which terms do not have UMLS codes. The coverage of UMLS codes for the terms is shown above the table (e.g., in Fig. 5, the UMLS coverage for atomic procedure terms in the top level is 94.02%). Users can also see the original text where the terms come from by clicking a row of the table. The original text will be displayed underneath the table as shown on the lower right pane.

Unlike Single Mode, Compositional Mode displays detailed information for the entire path from a selected node to the root when a node is selected (e.g., Fig. 8). The elementary concepts will be concatenated using "->" to indicate the path along the semantic tree. Displayed UMLS codes will be the ones that cover the entire (the most specific) meanings conveyed by the concatenated concepts. In this way, the coverage of UMLS codes for a specific pathological procedure, as well as for a specific semantic structure represented by a path from root to a selected node of the tree, will be easily investigated. For example, in Fig. 8, the node "region" in the left pane was selected. Because the Compositional Mode is chosen, the upper right pane displays the detailed information of the

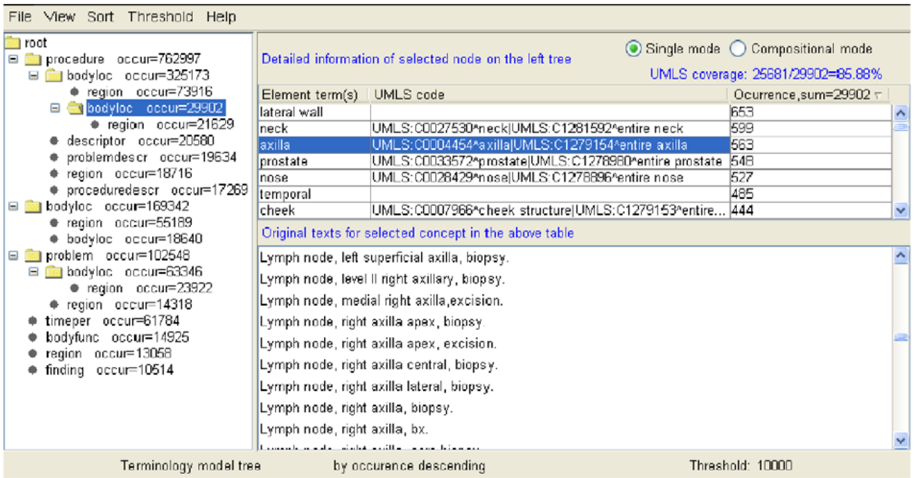

Fig. 7. A screen snapshot of TMviewer (Threshold = 10,000). In the left pane, the node "bodyloc" (body location) in the deepest level of the first tree is selected. The right pane displays the detailed information of the elementary concepts that belong to the semantic type "bodyloc." The Single Mode is selected and the table is sorted by occurrence. UMLS coverage for the terms of semantic type "location" in this level is 85.88%.

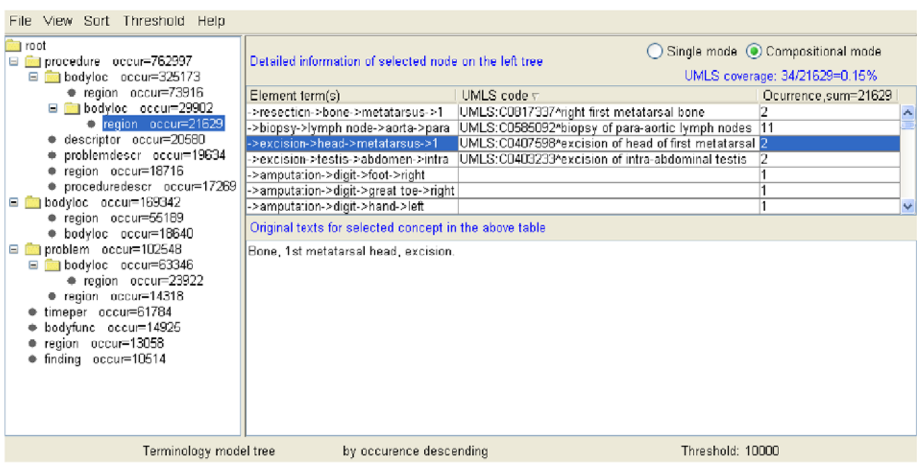

Fig. 8. A screen snapshot of TMviewer (Threshold = 10,000). In the left pane, the node "region" in the deepest level of the first tree is selected. The right pane displays the detailed information of the semantic path "procedure->bodyloc->bodyloc->region." It uses Compositional Mode, so the first column displays detailed information for the entire semantic structure along the path. The table is sorted by UMLS code. As a result, the compositional terms with or without a single UMLS code are grouped separately. The UMLS coverage for the compositional terms at this level is 0.15%.

semantic path "procedure->bodyloc->bodyloc->region." The first column in the table displays detailed terms for the entire semantic path. The UMLS coverage for this semantic path is 0.15%.

3.4. Terminology model generation

As a general medical language processor, MedLEE does not specify the relations between specific semantic types. Those edges in the generated semantic tree can be loosely regarded as "modify" relationships. Though the explicit relation is not stated, some relations could be assumed. Users can browse TMviewer and infer those relations based on the information provided by TMviewer and their own medical knowledge. For example, in Fig. 7, the semantic type *procedure* in the top level is connected to semantic type *body_loc* and both semantic types have large frequencies of occurrence. Thus, users may infer that there could be a probable relation between *procedure* and *body_loc*, or a user may want to analyze terms in display. If the user has enough medical knowledge, he (she) could infer that the relation between *procedure* and *body_loc* is *has_site*. Finally, the semantic relation *procedure*-(*has_site*)->*body_loc* can be established. Similarly, the semantic relation *body_loc*-(*has_region*)->*region* can be easily obtained from TMviewer. Depending on the granularity requirement of the model, connecting each of these semantic relations will result in a final terminology model.

4. Results

We collected a total of 888,357 pathology reports from 1991 to 2000, containing 358,260 unique terms of 1,406,785 occurrences which corresponded to pathology procedures. After removing non-semantic unique identifiers, 104,312 unique terms of 1,151,829 occurrences in total were identified. MedLEE successfully parsed 98,962 (94.9%) of these terms. UMLS Release 2005AA was used for this study. Among the 98,962 unique terms, only 5503 (5.6%) terms had a CUI for the complete term (e.g., "Bone marrow biopsy" has *C0005954*), 88,958 (89.9%) terms did not have a single CUI for the complete term but had separate CUIs for some of their components (e.g., there is no CUI for the complete term "gastric antral biopsy," so its components are assigned C0005558 for "biopsy," C0038351 for "stomach," and C0192420 for "biopsy of stomach"), and 4501 (4.5%) terms had no matching CUIs (e.g., "unstained slides"). The most frequent semantic types that appeared in the corpus are *procedure*, *bodyloc*, *region*, *problem*, and *descriptor*. Descriptions in MedLEE, examples, and percentages of these semantic types are shown in Table 1.

While this tool could be used to generate a complete and formal model for general pathology procedures by setting the threshold to zero, we built a simplified prototype model for demonstration purposes by using a threshold of 10,000 and reviewing the results provided by TMviewer. Our tool seeks to assist the model development process; however, the determination of elementary concepts and the development of rules to control the combination of concepts must still be performed by terminology developers.

We discovered three major patterns that physicians used in pathology reports. These patterns are shown in Fig. 9. Pattern 1, with the highest frequency of occurrence, uses *procedure* as the root. Other semantic types in this pattern include *descriptor*, *bodyloc*, and *region*. We rejected the pattern in which region directly modified the procedure because by manually reviewing the original terms, we found that the reason *region* (e.g., "junction" and "left") was a direct modifier for *procedure* instead of *bodyloc* was that MedLEE could not recognize a *bodyloc* if it occurred as an abbreviation or acronym (e.g., "G.E. Junction biopsy").

The other two patterns begin with *bodyloc* (Pattern 2) and *problem* (Pattern 3). After manually reviewing the original terms, we found that the reason these patterns do not contain any procedure is that clinicians often omit the procedure names and only enter the body location names. For example, the term "liver biopsy" should be entered, but only "liver" occurs. For Pattern 3, we found that clinicians might note down only associated health problems (e.g., "brain tumor," "endocervical polyp," and "breast mass") for a procedure instead of the name of the procedure.

Table 1
The most frequent semantic types that appeared in the corpus of pathology reports

Semantic type	Description in MedLEE	Example	% of each semantic type
Procedure	Terms denoting a therapeutic or diagnostic procedure	"biopsy"	39.5
Bodyloc	Terms denoting a well-defined area of the body or a body part	"liver"	32.5
Region	Terms denoting relative locations with a body location	"right"	13.1
Problem	Terms denoting a sign, symptom, disease, or syndrome	"ulcer"	5.3
Time period and date	Terms denoting temporal information	"19910212"	3.2
Descriptor	Terms qualifying a property of a body location/a measure of body (e.g., sat)/a finding/a function of body	"broad" in "broad area biopsy"	1.3
Problem_descriptor	Terms describing a problem or same as problem but occurs as modifier of problem	"degenerative" in "degenerative ganglionic cyst"	1.2
Procedure_descriptor	Terms describing a procedure or same as procedure but occurs as modifier of procedure	"digital" in "digital prostate biopsy"	1.1
Body_function	Terms denoting a body function	"respiration"	0.8
Finding	Terms signifying a normal or abnormal condition	"swelling," "alert"	0.5
Quantity	Terms denoting quantity information	"multiple"	0.3
Position	Terms denoting orientation	"diagonal"	0.3
Device	Terms denoting a medical device applied to patient	"chest tube"	0.2
Status	Terms relating to type of onset of finding or to time of onset, and other temporal information	"previous"	0.1
Others	For example, device descriptor	"prosthetic" in "prosthetic cuff"	0.6

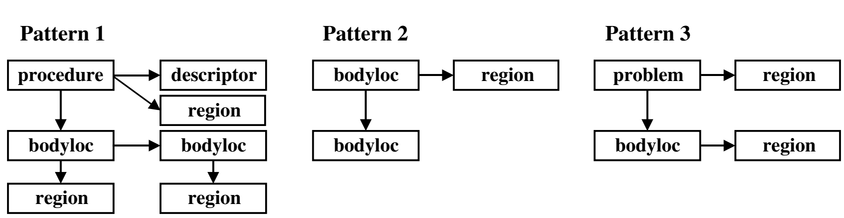

Fig. 9. Three major patterns for the pathology procedures that physicians used in the pathology reports found by applying MedLEE semantic types. Labeled rectangles are semantic types and arrows indicating the semantic relations of semantic types within the pattern.

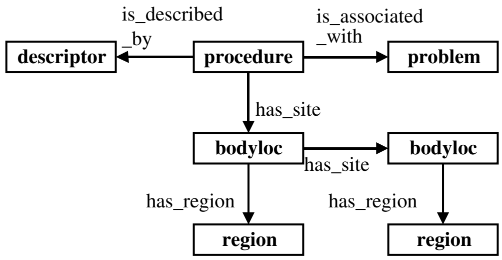

Fig. 10. A prototypical compositional concept-oriented terminology model for pathology procedures domain derived from the three patterns in Fig. 9. These three patterns were reorganized and merged due to reasons as stated in details in the text. *procedure* was added as the top-level to Pattern 2 and 3 in this figure, *bodyloc* was linked to *procedure* instead of *problem*, and then these two patterns were merged to the model in this figure. In the graphic representation, concepts are shown as labeled rectangles and relations are shown as labeled directional arcs.

Based on these three patterns and our medical knowledge, a prototypical and simplified terminology model for pathology procedure was established as shown in Fig. 10. It should be noted that this is just a simplified model for demonstration purpose and a complete model would be much more complex if the threshold was set to zero. Five semantic types including *procedure*, *problem*, *descriptor*, *bodyloc*, and *region* were identified, and associated relations including *is_associated_with*, *has_site*, *has_region*, and *is_described_by* were inferred and added to the model. The relationships among semantic types linked with those relations are illustrated in Fig. 10. As such, a procedure can be associated with a *problem* (such as "**breast mass** aspirate"), can be described by a *descriptor* (such as "**loop** excision of cervix"), and can have a site of *bodyloc* (such as "biopsy of **colon**"). The *bodyloc* can be further modified by a *region* or another nested *bodyloc*. For example, "4th toe bone resection" contains a semantic structure of procedure (resection) – bodyloc (bone) – bodyloc (toe) – region (4th).

5. Discussion

In this paper, we described an automated method for assisting the development of medical terminology models using NLP and information visualization techniques. Surgical pathology reports were selected as the corpus for developing a terminology model for pathology procedures, since this domain has a demand for developing terminology models to deal with compositional terms. Although MedLEE was originally designed for parsing radiology reports, subsequent modifications and extensions have made it a general NLP parser for a broader domain of medicine. An important point to note is that MedLEE's rules are general and have minimal constraints that restrict semantic relations in order to accommodate different domains. For example, modifiers of procedures, problems, devices, and descriptive findings are similar across different medical domains, and therefore are not specific to surgical pathology procedures.

Constructing a terminology model is a complex task. Besides data processing, it also requires determination of elementary concepts, rules that constrain the relationships among concepts (semantics), and rules that specify how the elementary concepts may be combined (syntax). In the following, first, we talk about the major advantages of our method in assisting terminology model discovery in general, and then discuss how it helps to address some of above issues by dividing the task into more trackable subtasks. We also report several interesting findings through the use of this tool. Lastly, we discuss some limitations of this method.

With traditional manual methods, collecting, aggregating, and formatting data are time-consuming. For example, if it takes a terminology model developer 30 s to read procedure terms in a pathology report, reviewing and collecting terms from a large corpus (e.g., 888,357 reports) could take more than 2 years assuming 10 h of work per day. Additionally, summarizing and finding patterns can be a labor-intensive and error-prone process since rare and possibly important terms and patterns might be overlooked. For example, in Fig. 6, a rare semantic structure, "procedure->bodyloc->region->region->locative" from the term "bladder, adjacent left floor, biopsy," has only one instance. In our system, a terminology developer can easily find rare terms like this by setting the threshold to zero. Considering the total number of instances ($n = 762{,}997$), without any assistance, this structure could be easily missed or its detection could be time-consuming. Our techniques can handle large text corpora and shorten the length of the early development steps by assisting data exploration and allowing model developers to focus on building the models. In our case, the whole process from corpus collection to preliminary modeling only took a few days. We believe that our system could also help to shorten the time needed to develop a complete terminology model.

In this study, we demonstrated our method in the pathology domain and used data from our institute. Another advantage of our method is that it is generalizable. The combination of NLP and visualization techniques can be easily applied to other data sources of the same domain or other clinical domains such as nursing or anatomy by simply loading the data into the system.

One important feature of our method is visualizing the coding of UMLS. MedLEE has a function for automated mapping of clinical documents to a comprehensive terminology thesaurus, the UMLS. The visualization tool provides two ways to view the UMLS coding. In the Single Mode, the coverage of UMLS coding for each atomic term can be assessed. In the Compositional Mode, the semantic path and the most specific UMLS encoding are presented. This is an important information source for users in terminology model development, which allows user to determine how information is incoded in its constituent vocabularies.

TMviewer provides a visualization environment which displays both merged semantic structures and the original terms along with their frequencies of occurrence. By letting users view the frequency of each term (by using Single Mode) and composite term (by using Compositional Mode), and their positions in the semantic structure, TMviewer may help users to decide what the elementary concepts should be. In addition, the semantic tree can be shown in detail and can also be easily pruned to show the prominent patterns by setting the frequency of occurrence threshold. While walking through the tree, users may get hints as to how the terms relate to each other, which relations are optional, which are obligatory and which are not allowed, and then decide what kinds of principles are needed to constrain the relationship and furthermore control the combination of terms.

Several interesting findings were discovered in this study using our method. First, as we know, SNOMED has a concept model for procedure (the SNOMED procedure hierarchy includes all clinical actions and health care services, such as surgical procedure, laboratory procedure, nursing procedure, and so on). It has a set of attributes, such as procedure site, procedure device, method, and specimen. Most of these attributes were present in the pathological procedures in our data, though the attribute names might be different. However, more elaborate information was revealed. For example, the phrase "EUS (Endoscopic ultrasonography) guided FNA (fine needle aspiration) of 2 cm pancreatic mass" contains a quantitative measure of the specimen and two procedure devices where one (EUS guided) functions as a descriptor of another (FNA). "Hysterectomy, right and left parametria multiple lymph nodes" contains an unspecified number for the quantity of lymph nodes in two body locations (right and left parametria). With respect to time, phrases may not only include absolute dates (e.g., 02/30/1995), but also relative time (e.g., "left breast taxol final day aspirate"). The sample findings listed above suggest that detailed information such as quantity and relative time might need to be considered for procedure terminology models. Second, the relations of these semantic types are often nested. For example, "skin, right eye lower lid outer lateral margin, biopsy" where biopsy has a site, but this site is not stated with a simple word. Rather, it is a combination of body locations and regions. We expect that the detailed semantic relationships among the terms and the patterns discovered by the tool from a large corpus will help build principles and rules needed to constrain the combinations in order to avoid redundant or inconsistent encoding. Third, we found that for many of the terms, *procedure* (e.g., biopsy) were missing and physicians only mentioned modifiers such as *bodyloc*, *device*, *problem*, and *finding*. This missing information causes incompleteness and vagueness, and affects automated data processing, indicating the need of formal models for encoding detailed clinical information.

We also realize several limitations in our study. First, for an NLP system to process free-text terms, the term has to exist in the lexicon of the NLP system. Otherwise, the element concept associated with the word will be lost. In our case, MedLEE failed to parse 5.1% of the terms. However, many of these terms were jargon used in the local hospital (e.g., "most recent three" which are three procedures changed from time to time), or contained no specific information (e.g., "unlabeled specimen," "slides"). Some failures were due to abbreviations and acronyms used by physicians (e.g., "G.E. Junction biopsy" in which G.E. means gastroesophageal, and others such as "O.S." and "O.D.," which are abbreviations for oculus sinister (left eye) and oculus dexter (right eye)), and truncated words such as "rec'd." The second limitation is that the semantic types that generated the models are limited by MedLEE's semantic types. Therefore, it is possible that some semantic types cannot be obtained by our method. Third, the semantic structure generated completely depends on MedLEE's semantic grammar. An incomplete grammar can cause the failure of capturing certain semantic structures. Future work will involve testing our method's generalizability to other clinical domains and conducting formal evaluation.

6. Conclusion

In this paper, we presented an automated high-throughput method for developing medical terminology models based on NLP and information visualization techniques. Surgical pathology reports were selected as the testing corpus for developing a pathology procedure concept model. The use of a general NLP processor, MedLEE, provides an automated method for extracting semantic structures from a free text corpus and sheds light on a new high-throughput method of medical terminology model development. Compared with traditional manual construction methods for terminology model development, this method can work on large text corpora in an automated fashion. Thus, the generated model could be more representative and complete. The use of information visualization techniques in our method reduces the burden of human developers by summarizing and visualizing the large quantity of semantic structures generated by NLP. This method, based on NLP and information visualization, may facilitate the modeling of medical terminologies for other domains.

Acknowledgment

This study is partially supported by the National Library of Medicine Grants R01 LM06274.

References

[1] Cimino J. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4–5):394–403.
[2] McDonald C. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4(3):213–21.
[3] Evans D, Cimino J, Hersh W, Huff S, Bell D. Toward a medical-concept representation language. The Canon Group. J Am Med Inform Assoc 1994;1(3):207–17.
[4] Campbell K, Das A, Musen M. A logical foundation for representation of clinical data. J Am Med Inform Assoc 1994;1(3):218–32.

[5] Spackman K, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp 2002:712–6.

[6] Friedman C, Cimino J, Johnson S. A schema for representing medical language applied to clinical radiology [published erratum appears in J Am Med Inform Assoc 1994;1(3):248]. J Am Med Inform Assoc 1994;1(3):233–48.

[7] Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The Canon Group's effort: working toward a merged model. J Am Med Inform Assoc 1995;2(1):4–18.

[8] Rocha RA, Huff SM. Development of a template model to represent the information content of chest radiology reports. Medinfo 2001;10(Pt. 1):251–5.

[9] Ozbolt J. Terminology standards for nursing: collaboration at the summit. J Am Med Inform Assoc 2000;7(6):517–22.

[10] Button PS, Androwich I, Mead CN, Zingo C, Konicek D, Campbell KE. Challenges in the development and testing of a reference terminology model for nursing interventions. Medinfo 2001;10(Pt. 1):176–80.

[11] Matney S, Dent C, Rocha RA. Development of a compositional terminology model for nursing orders. Int J Med Inform 2004;73(7–8):625–30.

[12] Rosse C, Mejino J. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478–500.

[13] Rosse CSL, Brinkley JF. The digital anatomist foundational model: principles for defining and structuring its concept domain. Proc AMIA Symp 1998:820–4.

[14] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Medinfo 2004;11:444–8.

[15] Mejino JJ, Agoncillo A, Rickard K, Rosse C. Representing complexity in part-whole relationships within the Foundational Model of Anatomy. AMIA Annu Symp Proc 2003:450–4.

[16] Michael J, Mejino J, Rosse C. The role of definitions in biomedical concept representation. Proc AMIA Symp 2001:463–7.

[17] Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, et al. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Int J Med Inform 2000;58–59:71–85.

[18] Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, et al. Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Stud Health Technol Inform 1999;68:901–5.

[19] Rossi Mori A, Galeazzi E, Consorti F. An ontological perspective on surgical procedures. Proc AMIA Annu Fall Symp 1996:115–9.

[20] Wagner J, Rogers J, Baud R, Scherrer J. Natural language generation of surgical procedures. Int J Med Inform 1999;53(2–3):175–92.

[21] Rector A, Nowlan W. The GALEN project. Comput Methods Programs Biomed 1994;45(1–2):75–8.

[22] Starren J, Johnson SB. Expressiveness of the breast imaging reporting and database system (BIRADS). Proc AMIA 1997;1:655–9.

[23] Bell D, Pattison-Gordon E, Greenes R. Experiments in concept modeling for radiographic image reports. J Am Med Inform Assoc 1994;1(3):249–62.

[24] Matney S, Dent C, Rocha RA. Development of a compositional terminology model for nursing orders. Int J Med Inform 2004;73(7–8):625–30.

[25] Bakken S, Warren JJ, Lundberg C, Casey A, Correia C, Konicek D, et al. An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED Clinical Terms. Int J Med Inform 2002;68(1–3):71–7.

[26] Bakken S, Warren J, Lundberg C, Casey A, Correia C, Konicek D, et al. An evaluation of the utility of the CEN categorical structure for nursing diagnoses as a terminology model for integrating nursing diagnosis concepts into SNOMED. Medinfo 2001;10(Pt. 1):151–5.

[27] Brown P, Price C. Semantic based concept differential retrieval and equivalence detection in clinical terms version 3 (Read Codes). Proc AMIA Symp 1999:27–31.

[28] Robinson D, Price C, Brown P. Clinical administration procedures in the Read Thesaurus: extending the ENV 1828 model to support regional terminology requirements. Proc AMIA Symp 1998.

[29] Cimino JJ. Terminology tools: state of the art and practical lessons. Methods Inf Med 2001;4:298–306.

[30] Baud RH, Lovis C, Ruch P, Rassinoux AM. A light knowledge model for linguistic applications. Proc AMIA Symp 2001:37–41.

[31] Ceusters W, Rogers J, Consorti F, Rossi-Mori A. Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN. Artif Intell Med 1999;15(1):5–23.

[32] Friedman C, Liu H, Shagina L. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. J Biomed Inform 2003;36(3):189–201.

[33] Liu H, Friedman C. A method for vocabulary development and visualization based on medical language processing and XML. Proc AMIA Symp 2000:502–6.

[34] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.

[35] Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:270–4.

[36] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74(8):890–5.

[37] Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inform Assoc 1999;6(1):76–87.

[38] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.

[39] Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, Clayton P. Unlocking clinical data form narrative reports. Ann Intern Med 1995;122(9):681–8.

[40] Lussier Y, Shagina L, Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp 2001:418–22.

[41] Cao H, Stetson P, Hripcsak G. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. J Biomed Inform 2003;36:99–105.

[42] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005:M1794.

[43] Mendonca E, Hass J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005;38(4):314–21.

[44] Wilcox A, Hripcsak G. The role of domain knowledge in automating medical text report classification. J Am Med Inform Assoc 2003:330–8.

[45] Xu H, Anderson K, Grann V, Friedman C. Facilitating cancer research using natural language processing of pathology. Medinfo 2004:565–72.

[46] Chuang J, Friedman C, Hripcsak G. A comparison of the Charlson comorbidities derived from medical language processing and administrative data. Proc AMIA Symp 2002:160–4.

[47] Hripcsak G, Austin J, Alderson P, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224(1):157–63.

[48] Bakken S, Hyun S, Friedman C, Johnson S. A comparison of semantic categories of the ISO reference terminology models for nursing and the MedLEE natural language processing system. Medinfo 2004:472–6.

[49] Friedman C, Liu H, Shagina L. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. J Biomed Inform 2003;36(3):189–201.

[50] UMLS knowledge sources, 2005AB Release Documentation. NLM, NIH.